Eksperiment Čarovnik iz Oza

Melita Hajdinjak, France Mihelič

Laboratorij za umetno zaznavanje, sisteme in kibernetiko,
Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
{melita.hajdinjak,france.mihelic}@fe.uni-lj.si

Povzetek

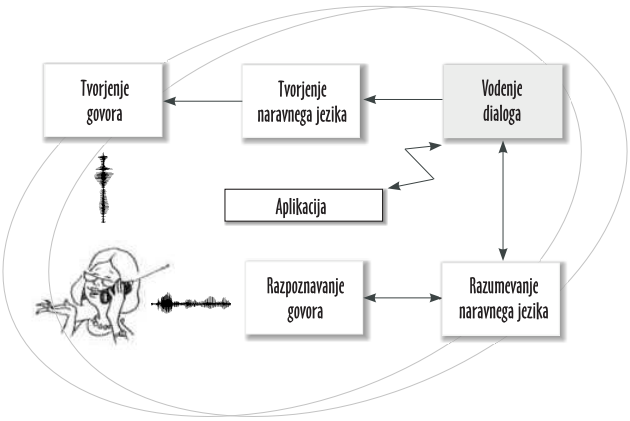
Opisali bomo izvajanje in podali rezultate eksperimenta *Čarovnik iz Oza*, s katerim s pomočjo človeka (čarovnika) simuliramo delovanje še nedokončanega sistema za dialog. Pri razvoju sistema za podajanje informacij o vremenu smo ta eksperiment uporabili dvakrat. Pri izvajanju prvega eksperimenta smo uporabili sistem *Čarovnik iz Oza*, katerega sestavni deli so bili: grafični vmesnik, zbirka tabel vremenskih podatkov, modul za tvorjenje naravnega jezika in modul za tvorjenje slovenskega govora. V drugem eksperimentu smo v sistem vključili še modul za vodenje dialoga, naloga čarovnika pa je ostala simulirati le delovanje modula za razpoznavanje govora in modula za razumevanje naravnega jezika.

Wizard-of-Oz experiment

We describe the conducted *Wizard-of-Oz* experiment where the developed dialogue system is partly simulated by a human wizard. While developing the spoken natural-language dialogue system for weather-information retrieval this experiment was used twice. In the first experiment, the *Wizard-of-Oz* system consisted of a graphical interface, a weather database, a natural-language generation module and a Slovenian text-to-speech module. In the second experiment the dialogue manager was added and the role of the wizard was to simulate speech recognition and natural-language understanding.

1. Uvod

Sistem za dialog ali govorni vmesnik imenujemo računalniški sistem, ki uporabniku omogoča, da z govorom dostopa do določenih aplikacij. Poznamo več vrst sistemov za dialog, najbolj razširjeni pa so *sistemi za podajanje informacij*, *sistemi za govorno upravljanje naprav*, *sistemi za interaktivne govorne odzive* in *sistemi za reševanje problemov*. Slikovit pregled najpogostejših sistemov za dialog, ki so ponavadi modularno zgrajeni (slika 1), je podan v (Krahmer, 2001).



Slika 1: Modularno zgrajen sistem za dialog

Prihodnost govornih vmesnikov temelji na potrebi po dostopu do informacij preko telefona, na potrebi po uporabi govora, ko so roke ali oči kako drugače zaposlene, ter kot pomoč invalidnim osebam. Sodobnejši sistemi za podajanje informacij preko telefona pa so usmerjeni na različne domene, npr. podajanje informacij o restavracijah (Jurafsky et al., 1994), o gledališčih (van der Hoeven, 1995), o potovanjih z železnico (Allen et al., 1995; Sturm et al., 1999), o letalskih prevozih (Ipšić et al., 1999; Stallard, 2000), o vremenskih napovedih (Zue et al., 2000) in prispeli elektronski pošti (Walker, 2000). Njihov cilj je strategije interakcije čimbolj približati naravnim strategijam, t.j. strategijam komunikacije človek – človek. To pomeni, da se je po eni strani potrebno omejiti le na najpogosteje zaznane naravne strategije, po drugi strani pa je potrebno razvijati in uporabljati najnovejše oz. najučinkovitejše tehnologije na področju obdelave govornega signala in razumevanja naravnega jezika.

2. Namen članka

Raziskovalci Laboratorija za umetno zaznavanje, sisteme in kibernetiko Fakultete za Elektrotehniko v Ljubljani ter raziskovalci Oddelka za računalništvo in informatiko Filozofske fakultete na Reki smo si zastavili cilj razviti sistem (Žibert et al., 2003), ki bo preko telefona v slovenskem in v hrvaškem jeziku podajal informacije o vremenu in vremenskih napovedih. Eden izmed razlogov za to odločitev je veliko število turističnih izmenjav med obema državama, drugi pa čedalje manjši delež prebivalcev obeh držav, ki govorijo in razumejo oba jezika. Sistem naj bi bil sposoben odgovarjati na vprašanja o vremenu in vremenski napovedi (splošna vremenska napoved, biovremenska napoved, temperatura, smer in hitrost vetra, zračni tlak, vidljivost, višina in vrsta snega, vzhod in zahod sonca, ipd.) za različne kraje in pokrajine v Sloveniji in na Hrvaškem ter za večja mesta ostalih evropskih držav.

Ker naj bi sistem omogočal hkratno komunikacijo v obeh jezikih, t.j. v slovenskem in v hrvaškem jeziku, bo uporabnikova izjava najprej potovala v modul za identifikacijo jezika, ki bo ugotovil, v katerem jeziku uporabnik govori. V skladu s to ugotovitvijo bo izjava posredovana v ustrezno zaporedje modulov za razpoznavanje govora (Martinčić-Ipšić et al., 2003) in razumevanje naravnega jezika, pri čemer bosta oba modula za razumevanje naravnega jezika pri pretvorbi izjave v njeno pomensko predstavitev upoštevala tudi trenutno stanje dialoga. Modul za vodenje dialoga bo (če bo potrebno), neodvisno od jezika, v zbirki vremenskih podatkov (Hajdinjak in Mihelič, 2002; Hajdinjak, 2004a) poiskal ustrezne podatke in jih, primerno strukturirane, poslal ustreznemu modulu za tvorjenje naravnega jezika. Ta bo to pomensko predstavitev odziva pretvoril v ustrezen (slovenski ali hrvaški) naravni jezik, ki bo nato še sintetiziran (Vesnicer, 2003), t.j. umetno pretvorjen v govor, in posredovan uporabniku preko telefonske linije.

Preden smo lahko pričeli z razvojem opisanega sistema, smo potrebovali relevantne podatke. Analiza dialogov človek – človek sicer predstavlja dobro osnovo za določitev nalog sistema in slovarja besed, ki naj jih sistem razume, ni pa preprosto ugotoviti, kateri vidiki te analize bodo ustrezali dialogom človek – računalnik (Smith in Gordon, 1997). Raziskovalci smo torej ujeti v začaran krog – za konstrukcijo sistema za dialog po eni strani potrebujemo značilnosti dialogov človek – računalnik, po drugi strani pa je, dokler sistem ne obstaja, nemogoče vedeti, kako bodo dialogi potekali.

Trenutno najboljša alternativa za zbiranje podatkov, ki odražajo jezik dialogov človek – računalnik, je tako imenovan *eksperiment Čarovnik iz Oza*. V teh eksperimentih so uporabniki prepričani, da se pogovarjajo z računalnikom, kar pa ni res. V resnici za računalnikom sedi človek (čarovnik), ki simulira delovanje sistema za dialog. V nekaterih primerih (Whittaker in Stenton, 1989; Eskenazi et al., 1999) čarovnik simulira celoten sistem, v drugih (Dahlbäck et al., 1993; Kim in Koo, 1997) pa le del sistema. Ključna je ugotovitev, da podatki, pridobljeni z eksperimentom Čarovnik iz Oza, bolj natančno odražajo jezik komunikacije človek – računalnik kot dialogi človek – človek (Whittaker in Stenton, 1989; Fraser in Gilbert, 1991; Dahlbäck et al., 1993). Glavni razlog za to je prilagajanje udeležencev dialoga jezikovnim sposobnostim sogovornika.

Eksperimentalni del razvoja sistema za dialog zato običajno poteka v treh korakih. V prvem koraku opravimo analizo dejanskih dialogov človek – človek. Na osnovi te analize konstruiramo enega ali celo več sistemov Čarovnik iz Oza, ki služijo kot ogrodje za izvedbo istoimenskega eksperimenta. Zadnji korak je izboljševanje in dopolnjevanje sistema za dialog s pomočjo podatkov dejanskih uporabnikov.

Pri razvoju sistema za podajanje informacij o vremenu, ki je predmet tega članka, smo prva dva koraka že naredili. Najprej smo vzpostavili osebno komunikacijo s *Hidrometeorološkim zavodom Agencije Republike Slovenije za okolje*, ki nudi telefonske pogovore z dežurnim prognostikom, na podlagi katerih so nam znali povedati, kaj dejanske uporabnike sploh zanima in kako po teh podatkih poizvedujejo. Te informacije so nam bile vodilo pri oblikovanju prvega sistema Čarovnik iz Oza (Hajdinjak in Mihelič, 2003a; Hajdinjak in Mihelič, 2003b). Kasneje, ko smo razvili še modul za vodenje dialoga, smo, z namenom vrednotenja delovanja tega modula, eksperiment še enkrat ponovili. V članku bomo poleg opisa izvajanja obeh eksperimentov Čarovnik iz Oza podali še rezultate in zaključke vrednotenja podatkov, pridobljenih v teh eksperimentih. Pri tem se bomo osredotočili na razvoj slovenskega dela sistema.

3. Prvi eksperiment Čarovnik iz Oza

Sistem Čarovnik iz Oza, ki preko telefona podaja informacije o vremenu in vremenski napovedi (Hajdinjak in Mihelič, 2003a), s katerim smo izvajali prvi eksperiment Čarovnik iz Oza, je modularno zasnovan, vsi moduli pa so povezani z grafičnim vmesnikom, za katerim je sedel človek (čarovnik) in simuliral del sistema za dialog. Sistem je udejanjen na osebnem računalniku z vgrajeno ISDN DIVA Server BRI-2M PCI kartico, s katero lahko komuniciramo preko posebnega računalniškega programa, ki omogoča vzpostavljanje in prevzemanje telefonskih pogovorov, poslušanje in snemanje pogovorov, kodiranje in dekodiranje ISDN formata zvočnih datotek, pošiljanje wave datotek po telefonski liniji, ipd.

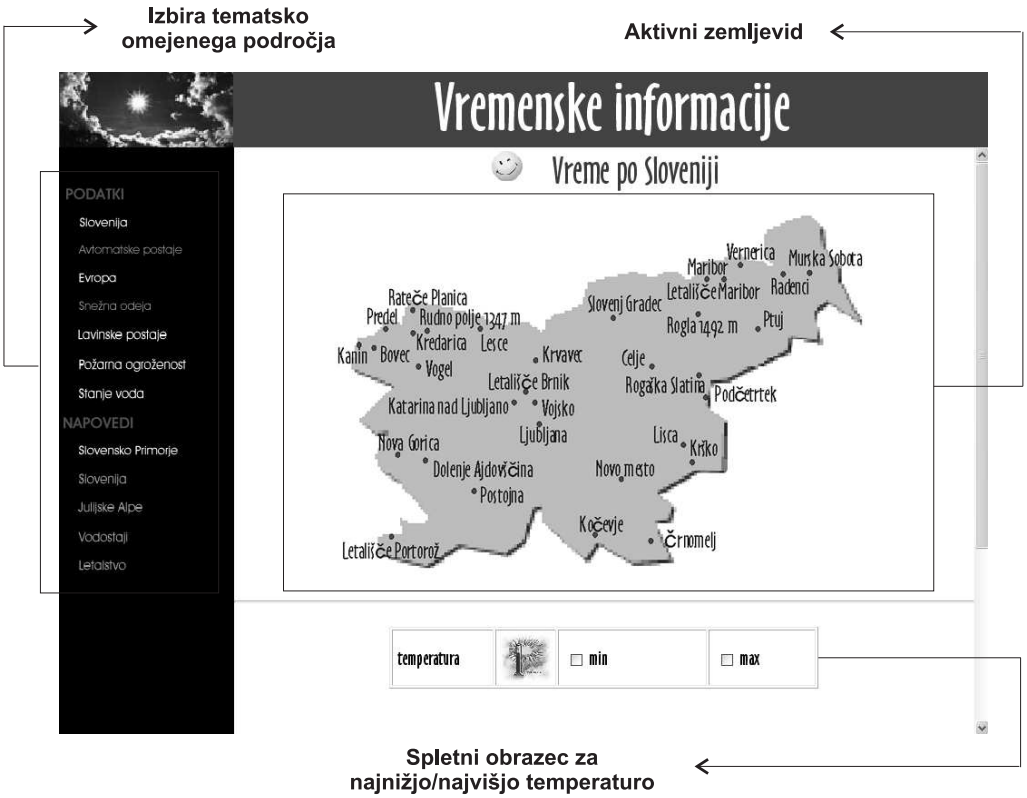
3.1. Zgradba prvega sistema Čarovnik iz Oza

Sistem Čarovnik iz Oza, uporabljen v prvem eksperimentu, sestavljajo:

- platforma za ISDN telefonijo,
- zbirka podatkovnih tabel, do katerih je dostopal čarovnik,
- grafični vmesnik, ki je čarovniku omogočal vodenje dialoga,
- modul za tvorjenje naravnega jezika in
- modul za tvorjenje slovenskega govora.

V zgodnjih fazah razvoja sistema za dialog je primerneje, če avtomatičnega razumevanja govora ne vključimo v sistem, saj v tem primeru ne bomo omejeni z naglasi, spolom in starostjo uporabnikov. V literaturi najdemo dva načina simuliranja razumevanja govora – napake razumevanja govora lahko sistematsko vključimo v besedilo, na katerega se čarovnik odziva (Pirker et al., 1999), ali pa simuliramo takorekoč popolno razumevanje govora (Dahlbäck et al., 1993; Kim in Koo, 1997; Eskenazi et al., 1999). Za slednji način smo se odločili tudi mi, čeprav se zavedamo, da so napake razumevanja govora izjemno pomemben vidik uporabnikove interakcije s sistemom. Za ta pristop smo se odločili, ker je napake razumevanja govora, še posebej pa napake razpoznavanja govora, zelo težko ustrezno simulirati. Naš čarovnik je zato poslušal uporabnika, brez da bi ga pri tem kakorkoli omejevali.

Eksperiment je potekal tako, da je čarovnik sedel za grafičnim vmesnikom, poslušal izjave uporabnika, v skladu s katerimi je po lastni presoji izbiral odzive na grafičnem vmesniku in s tem po potrebi dostopal do vremenske zbirke



Slika 2: Grafični vmesnik sistema Čarovnik iz Oza – primer delovnega okna

podatkovnih tabel (Hajdinjak in Mihelič, 2002). Odzivi so potovali v modul za tvorjenje naravnega jezika, katerega izhod (naravno besedilo) je potoval v modul za tvorjenje slovenskega govora, tega pa je sistem po telefonski liniji poslal do uporabnika. Naloga čarovnika v prvem eksperimentu Čarovnik iz Oza je torej bila igrati vlogo modulov za razpoznavanje govora in razumevanje naravnega jezika, s pomočjo grafičnega vmesnika pa tudi vlogo modula za vodenje dialoga. Za ta eksperiment smo se odločili v prvi fazi razvoja sistema za podajanje informacij o vremenu, da bi pridobili podatke, na osnovi katerih bi potem gradili omenjene module, ki jih je čarovnik v eksperimentu nadomeščal.

Grafični vmesnik, ki čarovniku omogoča vodenje dialoga, smo želeli oblikovati tako, da bi bili odzivi čarovnika po hitrosti in konsistentnosti čimbolj podobni odzivom računalnika. Zato smo ta grafični vmesnik zasnovali kot uporabniku (čarovniku) prijazno spletno aplikacijo, ki vsebuje spletne obrazce, aktivne slike, slikovna polja in spustne sezname ter omogoča izbiro krajev s pomočjo aktivnih zemljevidov, predvajanje vnaprej pripravljenih zvočnih datotek, uporabo bližnjic, ipd. Za obdelavo spletnih obrazcev uporabljamo programe, zapisane v skriptnem jeziku CGI, ki jih čarovnik izvaja s pritiski tipk na miški oz. tipkovnici. Primer delovnega okna je prikazan na sliki 2.

3.2. Izvajanje prvega eksperimenta

V prvem eksperimentu je sodelovalo 76 uporabnikov, in sicer 38 žensk in 38 moških. Pri izbiri smo pazili, da bi uporabniki predstavljali čimbolj reprezentativen vzorec (spol, starost, izobrazba, narečje, telefonska linija, okolje telefonskega pogovora). Povprečna starost uporabnikov je bila 34 let, povprečna izobrazba pa srednja šola. Zastopanih je bilo vseh osem osnovnih slovenskih narečij, pazili pa smo tudi na vrste telefonskih linij (analogna, ISDN, GSM) in na okolja (tiho okolje, šolski hodnik, ulica, prostor z več ljudmi, menza, vklopljen radio/TV-sprejemnik, ipd.), v katerih so se uporabniki v času telefonskega pogovora s sistemom Čarovnik iz Oza nahajali.

Pred začetkom eksperimenta smo udeležencem povedali, da se bodo pogovarjali z računalnikom, t.j. s sistemom za podajanje informacij o vremenu, in jim dali ustna navodila o splošni funkcionalnosti sistema. Ker smo želeli po eni strani pridobiti čimveč posnetkov, po drugi strani pa, kljub igranju vloge dejanskih uporabnikov, zagotoviti čimbolj realne in raznolike dialoge, smo vsakemu izmed udeležencev dodelili dve nalogi. Prva naloga je bila pridobiti določeno informacijo, druga pa je zajemala določen scenarij oz. situacijo, ki naj bi si jo udeleženec poskušal zamisliti. Primera takšnih nalog sta:

1. Poskušajte ugotoviti, ali v Ljubljani sije sonce.
2. Konec tedna bi radi šli v hribe. Kaj vas zanima?

in

1. Poskušajte ugotoviti, kakšne temperature lahko pričakujemo jutri.
2. Načrtujete izlet s kolesom. Kaj vas zanima?

Udeležencem smo omogočili tudi lastno izbiro vprašanj – povedali smo jim, da lahko po opravljenih nalogah dialog s sistemom nadaljujejo.

Po pogovoru s sistemom Čarovnik iz Oza smo uporabnike prosili, da izpolnijo vprašalnik, katerega prvi del je vseboval vprašanja o spolu, starosti, izobrazbi, zaposlitvi,

1. Ali ste sistem brez težav razumeli?
2. Ali vas je sistem razumel?
3. Ali ste brez težav prišli do odgovora na vaša vprašanja?
4. Ali je bila hitrost interakcije s sistemom primerna?
5. Ali ste na vsakem koraku dialoga vedeli, kaj morate povedati?
6. Ali se je sistem na vaše izjave odzival hitro (brez pojasnilnih vprašanj)?
7. Ali se je sistem obnašal tako, kot ste med dialogom od njega pričakovali?
8. Glede na vašo izkušnjo s sistemom za podajanje informacij o vremenu, ali boste (ko bo to mogoče) sistem poklicali, če vas bo zanimalo vreme?

Tabela 1: Del vprašalnika, s katerim so udeleženci eksperimenta ocenjevali sistem Čarovnik iz Oza

narečju, vrsti telefonske linije in okolju, v katerem so se v času dialoga s sistemom nahajali. Drugi del vprašalnika se je nanašal na dialog s sistemom in je, poleg vprašanja, ali so dobili odgovor na prvo nalogo, zajemal različne vidike njihove interakcije s sistemom (tabela 1). Vprašanja, s pomočjo katerih so udeleženci ocenjevali sistem Čarovnik iz Oza, smo povzeli po (Walker et al., 1997), sprašujejo pa po učinku modula za tvorjenje govora, učinku modula za razpoznavanje govora, težavnosti pridobivanja informacij, hitrosti interakcije, izkušenosti uporabnikov, ustreznosti odzivov sistema, pričakovanem obnašanju sistema in načrtovani rabi sistema v prihodnosti. Odgovori so bili podani z lestvico od 1 (nikakor se ne strinjam) do 5 (popolnoma se strinjam).

	WOZ1	WOZ2
Tvorjenje govora	4.42	4.29
Razpoznavanje govora	4.51	4.29
Pridobivanje informacij	4.27	3.74
Hitrost interakcije	3.94	3.76
Izkušenost uporabnikov	4.40	4.28
Ustreznost odzivov	4.23	3.76
Pričakovano obnašanje	4.31	4.04
Raba v prihodnosti	3.99	3.78

Tabela 2: Povprečne ocene sistema Čarovnik iz Oza v prvem (WOZ1) in v drugem (WOZ2) eksperimentu Čarovnik iz Oza

Povprečne ocene uporabnikov iz obeh eksperimentov so podane v tabeli 2. Najslabše je bila ocenjena hitrost interakcije s sistemom, in sicer 3.94, kar pa niti ni tako slabo, če vemo, da je bil povprečen čas čakanja na odziv sistema 5.57 sekund. Večji del tega časa je porabil čarovnik, da je posredoval izbran odziv. Vse ostale ocene, od katerih se večina vsaj delno nanaša na učinek čarovnika, ležijo med 3.99 in 4.51.

V eksperimentu smo opazili nekaj zanimivosti. Prva je ta, da je večina uporabnikov svoja vprašanja oblikovala zelo podobno, ponavadi *Zanima me ...*, *Rad/Rada bi vedel/vedela ...*, *Mi lahko (prosim) poveste ...*, *Ali mi lahko poveste ...*, ipd. Druga zanimivost je prilagajanje uporabnikovega vedenja na pričakovane jezikovne sposobnosti sistema. V našem eksperimentu smo ugotovili, da je bilo prvo vprašanje, ki so ga uporabniki zastavili, ponavadi veliko daljše in veliko manj jedrnato kot vprašanja, ki so sledila. Primeri dolgih in nejedrnatih začetnih vprašanj so na primer:

⇝ Jaz moram danes na Primorsko, pa me je malo strah burje. Zdaj me pa zanima, kje začenja pihati, a na vetrišču ali šele s Svete gore navzdol proti Gorici.

⇝ Glejte. Trenutno se peljem z Jesenic proti Ljubljani. Razmere na cesti so obupne. Vseskozi piha močan veter s smeri vzhoda. Zdaj me pa zanima, kakšen veter piha v Ljubljani. Se pravi, moč pa smer, od kod piha.

V primerih, ko je sistem uporabnike prosil, naj vprašanje ponovijo, pa je njihov govor postal bolj artikuliran, počasnejši in/ali glasnejši. Že Zoltan-Ford (1991) je prišel do podobnih ugotovitev, namreč da jezikovne sposobnosti sistema za dialog opazno vplivajo na vedenje uporabnikov. Cohen in Oviatt (1995) pa trdita, da bodo uporabniki svoje besedišče zelo verjetno prilagodili sistemovemu, če bodo v njem zaznali omejitve.

Tretja zanimivost, ki smo jo opazili v prvem eksperimentu Čarovnik iz Oza, so smeh, šale in duhovite opazke, ki so jih nekateri uporabniki delali na račun sistema, medtem ko je čarovnik posredoval svoj odziv. Primeri takih opazk so:

⇝ Kakšen glas ... Grozno ...

⇝ Išče po računalniku ...

⇝ Razmišlja ...

Takšne reakcije in stranske opazke so najbrž bolj verjetne v testiranjih sistemov za dialog kot pa v komercialnih sistemih, saj v takih testiranjih uporabniki ponavadi igrajo le vloge dejanskih uporabnikov, niso pod časovnimi omejitvami in ne odgovarjajo za stroške telefonskega pogovora.

Primer dialoga med uporabnikom in sistemom Čarovnik iz Oza je podan v tabeli 3.

Zbirka spontanega govora, ki smo jo pridobili v prvem eksperimentu Čarovnik iz Oza, smo poimenovali SSSQ (Hajdinjak in Mihelič, 2003b). Kratica SSSQ se nanaša na angleško besedno zvezo *Slovenian Spontaneous Speech Queries*. Zbirko smo označili z orodjem TRANSCRIBER (Barras et al., 2001), pri čemer smo poleg oznak besed uporabili še posebne oznake za narečne besede, premore in negovorne zvoke (vdih, izdih, vzdihljaj, tlesk, smeh, pih,

SISTEM:	*Dober dan. Pogovarjate se s sistemom za podajanje informacij o vremenu. Kakšno je vaše vprašanje?*
UPORABNIK:	Mi lahko prosim poveste, kakšna bo danes vidljivost na Štajerskem?
SISTEM:	*Počakajte trenutek, prosim ...* [čarovnik izbira svoj odziv]
	Na Štajerskem bo danes podnevi vidljivost nad 10 km. Imate še kakšno vprašanje?
UPORABNIK:	A bo kaj snežilo?
SISTEM:	*Kateri kraj vas zanima?*
UPORABNIK:	Štajerska.
SISTEM:	*Počakajte trenutek, prosim ...* [čarovnik izbira svoj odziv]
	Ne, na Štajerskem danes podnevi – razjasnitve. Imate še kakšno vprašanje?
UPORABNIK:	Kakšno pa bo vreme v prihodnjih dneh na Poljskem?
SISTEM:	*Imam podatke za danes.*
UPORABNIK:	Torej, kako bo danes na Poljskem?
SISTEM:	*Počakajte trenutek, prosim ...* [čarovnik izbira svoj odziv]
	V Varšavi na Poljskem je oblačno, temperatura zraka je -6 stopinj celzija. Imate še kakšno vprašanje?
UPORABNIK:	A je kaj snega na Poljskem?
SISTEM:	*Te informacije ne ponujam. Imate morda kakšno drugo vprašanje?*
UPORABNIK:	Ne, najlepša hvala. Nasvidenje.
SISTEM:	*Hvala lepa za sodelovanje. Nasvidenje.*

Tabela 3: Primer dialoga med uporabnikom in sistemom Čarovnik iz Oza

kašelj in različne zvoke, ki nastajajo pri obotavljanju oz. razmišljanju).

4. Drugi eksperiment Čarovnik iz Oza

Ko smo razvili modul za vodenje dialoga (Hajdinjak, 2004a; Hajdinjak in Mihelič, 2004b) in ga vgradili v sistem, smo eksperiment Čarovnik iz Oza ponovili. Cilj in namen drugega eksperimenta je bil pridobiti podatke, ki bi poleg podatkov iz prvega eksperimenta tvorili osnovo za evalvacijo modula za vodenje dialoga.

Naloga čarovnika v drugem eksperimentu Čarovnik iz Oza je, v primerjavi s prvim eksperimentom, bila simulirati razpoznavanje govora in razumevanje naravnega jezika, ne pa tudi igrati vloge modula za vodenje dialoga. V tem eksperimentu je čarovnik sedel pred vmesnikom modula za vodenje dialoga in preko tipkovnice vnašal pomensko predstavitev uporabnikove izjave, ki bo v končnem sistemu izhod iz modula za razumevanje naravnega jezika. Vse nadaljnje delo (vodenje dialoga, iskanje podatkov, tvorjenje naravnega jezika, tvorjenje govora) je opravljal sistem.

V drugem eksperimentu je sodelovalo 68 uporabnikov, 29 žensk in 39 moških, katerih povprečna starost je bila 32 let. Čeprav je 17 izmed njih že sodelovalo v prvem eksperimentu Čarovnik iz Oza in so zato bili malo bolj izkušeni od ostalih, med vedenjem obeh skupin nismo opazili nobene razlike.

Uporabnike (tudi tiste, ki so že sodelovali v prvem eksperimentu) smo, tako kot v prvem eksperimentu, prepričali, da se bodo pogovarjali s sistemom za podajanje informacij o vremenu, in jim dali ustna navodila o splošni funkcionalnosti sistema. Vsakemu izmed udeležencev smo dodelili dve nalogi, podobno kot v prvem eksperimentu, od katerih je bila prva pridobiti določeno informacijo, druga pa scenarij oz. situacija, ki naj bi si jo udeleženec poskušal zamisliti. Po končanem pogovoru s sistemom smo uporabnike prosili, da izpolnijo enak vprašalnik kot v prvem eksperimentu. Del vprašalnika, ki zajema vprašanja, nanašajoča se na ocenjevanje sistema, je podan v tabeli 1.

Povprečne ocene udeležencev drugega eksperimenta Čarovnik iz Oza so podane v tabeli 2. Povprečni čas, ki ga je čarovnik porabil, da je modulu za vodenje dialoga posredoval pomensko predstavitev uporabnikove izjave, ki je sprožila odziv sistema, je sedaj znašal 6.61 sekund. Zanimivo je, da so udeleženci drugega eksperimenta slabše ocenili prav vse vidike svoje interakcije s sistemom, tudi učinek modulov za tvorjenje in razpoznavanje govora, ki sta v obeh eksperimentih ostala enaka. Glede na to, da so tako storili tudi tisti, ki so sodelovali v obeh eksperimentih Čarovnik iz Oza, razloga ne moremo pripisati večji kritičnosti udeležencev. Zelo verjetno pa so slabše ocene posledica manjšega zadovoljstva z nekaterimi posameznimi vidiki učinkovitosti sistema. Najbolj opazna razlika je v ocenah težavnosti pridobivanja informacij (4.27 oz. 3.74) in ustreznosti odzivov sistema (4.23 oz. 3.76), pri čemer je treba poudariti, da so bile slabše ocene nekaterih vidikov učinkovitosti sistema vsekakor pričakovane, saj bi v nasprotnem primeru modul za vodenje dialoga v drugem eksperimentu bolje opravljal svojo nalogo kot človeški operater v prvem eksperimentu.

Zbirko spontanega govora, pridobljeno v drugem eksperimentu Čarovnik iz Oza, smo iz podobnega razloga kot v prvem eksperimentu poimenovali SSSQ2 in jo označili v skladu z oznakami zbirke SSSQ. Drugi eksperiment so zaznamovali predvsem daljši dialogi (Hajdinjak, 2004a), in sicer tako po številu izjav kot tudi po časovnem trajanju. To je v glavnem posledica v drugem eksperimentu izvajane strategije vodenja dialoga (Hajdinjak, 2004a; Hajdinjak in Mihelič, 2004b), ki se od strategije čarovnika v prvem eksperimentu razlikuje predvsem po prošnjah sistema za potrditev posredovanih informacij. Te potrditve, zavrnitve oz. popravki so zaradi svoje jedrnatosti tudi glavni razlog v povprečju krajših izjav, saj so velikokrat

sestavljeni le iz ene ali dveh besed.

Da bi vsaj delno odpravili etično vprašljivost eksperimentov Čarovnik iz Oza, smo po obeh eksperimentih udeležencem povedali, kaj smo počeli, zakaj smo tako ravnali, in jih vprašali za dovoljenje uporabe pridobljenih podatkov v raziskovalne namene. Prav vsi so pokazali razumevanje in odobravanje teh eksperimentov, dovolili pa so nam tudi uporabo posnetkov in drugih podatkov, ki smo jih v eksperimentih pridobili.

5. Evalvacija sistemov iz obeh eksperimentov Čarovnik iz Oza

Za evalvacijo sistemov iz obeh eksperimentov smo uporabili potencialno splošno metodologijo evalvacije sistemov za dialog, namreč ogrodje PARADISE (PARAdigm for DIalogue System Evaluation) (Walker et al., 1997), ki omogoča izpeljavo ocene učinkovitosti sistema kot uteženo linearno kombinacijo od domene odvisnih *parametrov uspešnosti naloge* in *cen dialoga*. Model učinkovitosti sistema, ki ga zajema ogrodje PARADISE, trdi, da lahko funkcijo učinkovitosti sistema določimo s pomočjo *multiple linearne regresije* (MLR) z zadovoljstvom uporabnikov kot neodvisno spremenljivko ter parametri uspešnosti naloge in cen dialoga (*parametri učinkovitosti dialoga* in *parametri kvalitete dialoga*) kot neodvisnimi spremenljivkami.

5.1. Parametri modela učinkovitosti sistema

V obeh eksperimentih so uporabniki ocenili svoje zadovoljstvo tako, da so podali stopnjo strinjanja z izjavami o obnašanju oz. učinkovitosti sistema (tabela 1). Splošno **zadovoljstvo uporabnika** smo dobili tako, da smo vse točke te ankete sešteli. Vrednosti parametrov, ki smo jih uporabili kot vrednosti odvisne spremenljivke v MLR modelu učinkovitosti sistema, zato ležijo med 8 in 40. Del parametrov uspešnosti naloge in cen dialoga, ki v MLR modelu učinkovitosti predstavljajo neodvisne spremenljivke, smo določili samodejno, del pa smo ročno označili. V prvem eksperimentu smo tako določili parametra uspešnosti naloge

- **Kappa koeficient** (κ) in
- **izpolnitev naloge** (Comp), mnenje uporabnika o izpolnitvi prve naloge, ki smo mu jo v eksperimentu zastavili,

parameter učinkovitosti dialoga

- **povprečni čas dialoga** (MET), povprečni čas trajanja dialoga brez vštetega trajanja odzivov sistema,

in parametre kvalitete dialoga

- **povprečni čas čakanja na odziv** (MRT), povprečni čas, ki ga sistem porabi, preden se odzove,
- **delež nepodanih informacij** (NPR), delež začetnih potez uporabnika, na katere sistem ne poda relevantnega odgovora,
- **delež zavrnitev** (RR), delež izjav sistema, s katerimi uporabnika prosi, naj ponovi zadnjo izjavo,
- **delež nudenja pomoči** (HMR), delež izjav sistema, s katerimi uporabniku pomaga nadaljevati dialog,
- **delež negativnih odgovorov** (NDR), delež potez, s katerimi sistem sporoča, da nima zahtevanega podatka in pri tem uporabnika ne usmerja k izbiri relevantnih, dosegljivih podatkov,
- **delež izbire relevantnih podatkov** (RDR), delež potez sistema, ki uporabnika usmerjajo k izbiri relevantnih, dosegljivih podatkov, in
- **delež neprimernih iniciativ** (UIR), delež začetnih potez uporabnika, katerih vsebina ne ustreza domeni sistema.

V drugem eksperimentu Čarovnik iz Oza smo definirali dva dodatna parametra kvalitete dialoga, namreč

- **delež neiniciativnih potez** (NIR), delež potez uporabnika, ki ne spadajo med začetne poteze, in
- **delež preverjanj** (CR).

Srednje vrednosti vseh parametrov dialogov, pridobljenih v obeh eksperimentih Čarovnik iz Oza, so podane v tabeli 4.

5.2. Rezultati multiple linearne regresije

Pri izpeljavi funkcije učinkovitosti sistema z metodo multiple linearne regresije smo ugotovili, da so na zadovoljstvo uporabnikov v prvem eksperimentu najbolj vplivali **delež nudenja pomoči** (HMR), **delež nepodanih informacij** (NPR), **izpolnitev naloge** (Comp), **povprečni čas čakanja na odziv** (MRT) in **delež zavrnitev** (RR). Na zadovoljstvo uporabnikov v drugem eksperimentu Čarovnik iz Oza pa so najbolj vplivali **delež preverjanj** (CR), **Kappa koeficient** (κ), **povprečni čas dialoga** (MET), **delež nepodanih informacij** (NPR) in **izpolnitev naloge** (Comp). Natančen opis vrednotenja in izpeljave obeh funkcij učinkovitosti je podan v (Hajdinjak, 2004a).

Ker edino razliko sistemov, uporabljenih v teh eksperimentih, predstavlja način vodenja dialoga (prvič je to nalogo s pomočjo grafičnega vmesnika opravljal čarovnik, drugič pa modul za vodenje dialoga), razlike med obema funkcijama učinkovitosti sistema ponazarjajo učinek modula za vodenje dialoga.

V evalvaciji sistemov za branje elektronske pošte (Walker et al., 1998) je bilo ugotovljeno, da **izpolnitev naloge** močneje vpliva na **zadovoljstvo uporabnika** kot pa **Kappa koeficient**. Razlog, ki ga navajajo, je ta, da naj bi uporabniki velikokrat drugače dojemali delovanje sistema, kot ga podaja **Kappa koeficient**. V naših eksperimentih je bil **Kappa koeficient** takorekoč odvisen le od čarovnika, ki je nadomeščal razumevanje govora, parameter **izpolnitev naloge** pa se je nanašal le na prvo nalogo, ki smo jo uporabniku zastavili, kar je najverjetneje razlog, zakaj sami nismo prišli do podobnega zaključka. Po eni strani sta bila v naših eksperimentih **Kappa koeficient** in **izpolnitev naloge** nekorelirana (0.00 v prvem in 0.05 v drugem eksperimentu), po drugi strani pa je v drugem eksperimentu

	WOZ1	WOZ2
Kappa koeficient (κ)	0.94	0.98
izpolnitev naloge (Comp)	0.97	0.96
povprečni čas dialoga (MET)	15.41 s	22.03 s
povprečni čas čakanja na odziv (MRT)	5.57 s	6.61 s
delež nepodanih informacij (NPR)	0.31	0.28
delež zavrnitev (RR)	0.01	0.03
delež nudenja pomoči (HMR)	0.01	0.06
delež negativnih odgovorov (NDR)	0.08	0.07
delež izbire relevantnih podatkov (RDR)	0.06	0.20
delež neprimernih iniciativ (UIR)	0.13	0.04
delež neiniciativnih potez (NIR)	-	0.50
delež preverjanj (CR)	-	0.20
zadovoljstvo uporabnika (US)	34.08	31.96

Tabela 4: *Srednje vrednosti parametrov v prvem (*WOZ1*) in v drugem (*WOZ2*) eksperimentu Čarovnik iz Oza*

Kappa koeficient celo močneje vplival na **zadovoljstvo uporabnikov**.

Parametri, ki so pomembno vplivali na učinkovitost sistema, pri tem pa niso bili podvrženi čarovniku, so **delež nudenja pomoči** in **delež nepodanih informacij** v prvem ter **delež preverjanj** in **delež nepodanih informacij** v drugem eksperimentu. Vrednost parametra **delež nudenja pomoči** je odvisna od tega, kako se uporabnik v dialogu obnaša, to pa je spet odvisno od stopnje prijaznosti in sodelovanja, ki jo sistem pri vodenju dialoga nudi. Vsekakor, edini cilj sistema za dialog ne sme biti le uspešen zaključek naloge, ampak tudi sposobnost prevzeti iniciativo in nuditi pomoč, ko jo uporabnik potrebuje. Ker nekateri novi uporabniki sistema, ki se niso sposobni hitro prilagajati, pri vodenju dialoga pogosto potrebujejo pomoč, vpliva parametra **delež nudenja pomoči** na učinkovitost sistema ni mogoče odpraviti. Tudi vpliva parametra **delež preverjanj** ni mogoče odpraviti, saj razumevanje govora ponavadi predstavlja najtežavnejši del, če ne celo oviro do učinkovitosti sistema za dialog. Torej, zadovoljstvo uporabnikov lahko pomembno povečamo le z zmanjšanjem vpliva parametra **delež nepodanih informacij**, t.j. deleža začetnih potez uporabnika, na katere sistem ne poda relevantnega odgovora. Zmanjšanje vrednosti tega parametra lahko dosežemo tako, da sistemu ne dovolimo odziva, preden se ni zares prepričal, da ne dostopa do nobenega relevantnega podatka. Sistem naj torej uporabnika čimbolj fleksibilno usmerja k izbiri dosegljivih, relevantnih podatkov.

6. Zaključek

V članku smo opisali izvajanje in podali rezultate dveh eksperimentov Čarovnik iz Oza, ki so nam na eni strani služili za pridobivanje podatkov, potrebnih za razvoj sistema za podajanje informacij o vremenu, na drugi strani pa tudi za vrednotenje modula za vodenje dialoga.

Oba eksperimenta Čarovnik iz Oza sta pokazala, da na zadovoljstvo uporabnikov pomembno vpliva stopnja fleksibilnosti pri usmerjanju uporabnika k izbiri dosegljivih, relevantnih podatkov. To pa pomeni, da je predstavitev znanja v sistemu za dialog izrednega pomena. Čeprav smo v drugem eksperimentu že uporabili predstavitev znanja (Hajdinjak, 2004a), ki temelji na intuicionistični modalni logiki in v veliki meri omogoča ponujanje relevantnih, dosegljivih podatkov, je za izboljšanje zadovoljstva uporabnikov takorekoč nujno razmišljati o morebitnih dodatnih razširitvah oz. izboljšavah.

7. Literatura

J.F. Allen, L.K. Schubert, G. Ferguson, P. Heeman, C.-H. Hwang, T. Kato, M. Light, N.G. Martin, B.W. Miller, M. Poesio in D.R. Traum. 1995. The trains project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI*, 7:7–48.

C. Barras, E. Geoffrois, Z. Wu in M. Liberman. 2001. Transcriber: use of a tool for assisting speech corpora production. *Speech Communication: Special issue on Speech Annotation and Corpus Tools*, 33(1–2):5–22.

P. Cohen in H. Levesque. 1990. Rational Interaction as the Basis for Communication. V: *Intentions in Communication, MIT Press, Cambridge*, str. 221–255.

P. R. Cohen in S. L. Oviatt. 1995. The Role of Voice Input for Human-Machine Communication. V: *Proceedings of the National Academy of Sciences, ZDA*, zv. 92(22), str. 9921–9927.

N. Dahlbäck, A. Jönsson in L. Ahrenberg. 1993. Wizard of Oz studies: why and how. V: *Proceedings of the international workshop on Intelligent user interfaces, Orlando, ZDA*, str. 193–200.

M. Eskenazi, A. Rudnicky, K. Gregory, P. Constantinides, R. Brennan, C. Bennett in J. Allen. 1999. Data collection and processing in the carnegie mellon communicator. V: *Proceedings of the 6th European Conference on Speech Communication and Technology, Budimpešta, Madžarska*, str. 2695–2698.

N. M. Fraser in G. N. Gilbert. 1991. Simulating speech systems. *Computer Speech and Language*, 5(1):81–99.

M. Hajdinjak in F. Mihelič. 2002. Semantična analiza vremenskih napovedi. V: *Zbornik B 5. mednarodne multikonference Informacijska družba IS'2002, Ljubljana, Slovenija*, str. 10–13.

M. Hajdinjak in F. Mihelič. 2003a. Wizard of oz experiments. V: *Proceedings of the IEEE Region 8 EUROCON 2003 : computer as a tool, Ljubljana, Slovenija*, zvezek 2, str. 112–116.

M. Hajdinjak in F. Mihelič. 2003b. The wizard of oz system for weather information retrieval. *Lecture notes in computer science, Lecture notes in artificial intelligence, Text, speech and dialogue : 6th International Conference, České Budějovice, Češka*, 2807:400–405.

M. Hajdinjak. 2004a. *Vodenje dialoga med človekom in računalnikom v naravnem jeziku*. Magistrsko delo, Fakulteta za elektrotehniko, Univerza v Ljubljani, Ljubljana.

M. Hajdinjak in F. Mihelič. 2004b. Information-providing dialogue management. *Lecture notes in computer science, Lecture notes in artificial intelligence, Text, speech and dialogue : 7th International Conference, Brno, Češka*.

G. van der Hoeven, J. Andernach, S. van der Burgt, G. J. Kruijff, A. Nijholt, J. Schaake in F. de Jong. 1995. Schisma: A natural language accessible theatre information and booking system. V: *Proceedings of the 1st International Workshop on Applications of Natural Language to Data Bases, Versailles, Francija*, str. 271–285.

I. Ipšić, F. Mihelič, S. Dobrišek, J. Gros in N. Pavešić. 1999. A slovenian spoken dialog system for air flight inquires. V: *Proceedings of the 6th European Conference on Speech Communication and Technology, Budimpešta, Madžarska*, str. 2659–2662.

D. Jurafsky, C. Wooters, G. Tajchman, J. Segal, A. Stolcke, E. Fosler in N. Morgan. 1994. The berkeley restaurant project. V: *Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japonska*, str. 2139–2142.

W. Kim in M.-W. Koo. 1997. A korean speech corpus for train ticket reservation aid system based on speech recognition. V: *Proceedings of the 5th European Conference on Speech Communication and Technology, Rodos, Grčija*, str. 1723–1726.

E. J. Krahmer. 2001. The science and art of voice interfaces, philips research report. Tehnično poročilo, Philips, Eindhoven, Nizozemska.

S. Martinčić-Ipšić, J. Žibert, I. Ipšić, F. Mihelič in N. Pavešić. 2003. Bilingual speech recognition for a weather information retrieval dialogue system. *Lecture notes in computer science, Lecture notes in artificial intelligence, Text, speech and dialogue : 6th International Conference, České Budějovice, Češka*, 2807:380–387.

H. Pirker, G. Loderer in H. Trost. 1999. Thus spoke the user to the wizard. V: *Proceedings of the 6th European Conference on Speech Communication and Technology, Budimpešta, Madžarska*, zvezek 3, str. 1171–1174.

R. W. Smith in S. A. Gordon. 1997. Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue. *Computational Linguistics*, 23(1):141–168.

D. Stallard. 2000. Talk'n'travel: A conversational system for air travel planning. V: *Proceedings of the Association for Computational Linguistics Sixth Applied Natural Language Processing Conference*, str. 68–75.

J. Sturm, E. den Os in L. Boves. 1999. Dialogue management in the dutch arise train timetable information system. V: *Proceedings of the Sixth European Conference on Speech Communication and Technology*, str. 1419–1422.

B. Vesnicer. 2003. *Umetno tvorjenje govora z uporabo prikritih Markovovih modelov*. Magistrsko delo, Fakulteta za elektrotehniko, Univerza v Ljubljani, Ljubljana.

M. A. Walker, D. Litman, C. A. Kamm in A. Abella. 1997. Paradise: A general framework for evaluating spoken dialogue agents. V: *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, Madrid, Španija*, str. 271–280.

M. A. Walker, D. J. Litman, C. A. Kamm in A. Abella. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech and Language*, 12(3):317–347.

M.A. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

S. Whittaker in P. Stenton. 1989. User studies and the design of natural language systems. V: *Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics, Manchester, Anglija*, str. 116–123.

E. Zoltan-Ford. 1991. How to get people to say and type what computers can understand. *Journal of Man-Machine Studies*, 34:527–547.

V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen in L. Hetherington. 2000. Jupiter: A telephone based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, 8(1):85–96.

J. Žibert, S. Martinčić-Ipšić, M. Hajdinjak, I. Ipšić in F. Mihelič. 2003. Development of a bilingual spoken dialog system for weather information retrieval. V: *Proceedings of the 8th European Conference on Speech Communication and Technology, Ženeva, Švica*, str. 1917–1920.